Evaluate a security agent on your own findings

Trial record

[ ] Job and agent version: model ID and date, effort setting, prompt version, tool list
[ ] Test set: number of findings, where drawn from, who labeled them, date labeled, which findings are held back
[ ] Thresholds written before the run (copy them here with the date)
[ ] Verification method: the checks run in a clean copy and who ran them
[ ] Per finding: agent decision, reference answer, each check passed or failed
[ ] Wrong-but-confident cases listed in full
[ ] Blocked runs and retries, counted separately
[ ] Reviewer minutes per accepted change and per rejected one
[ ] Tokens used and cost per finding
[ ] Stop conditions that fired; any change attempted outside scope
[ ] Interval for the main rate (state the method)
[ ] Decision: widen, hold or stop, with the name of the person who decided and the date

Agentic AI in Cybersecurity, Evaluate a security agent on your own findings, edition 19 Nov 2026, https://agenticaiincybersecurity.com/guides/evaluate-a-security-agent-on-your-findings/. License CC BY 4.0.